# ICE: Intervention-Consistent Explanation Evaluation with Statistical Grounding for LLMs

Anonymous ACL submission

## Abstract

Evaluating explanation faithfulness—whether explanations reflect a model's true reasoning—remains challenging. Benchmarks like ERASER use single intervention strategies without statistical rigor, making it hard to separate genuine faithfulness from noise. We introduce ICE (Intervention-Consistent Explanation), a framework addressing these gaps through randomization tests with win rate metrics and bootstrap confidence intervals.

We evaluate 7 LLMs across 4 datasets and 4 languages using native sentiment data, comparing attention and gradient attribution. Key findings: (1) attention beats gradient on short text (+10–20%) but both converge on long text; (2) faithfulness and human plausibility are orthogonal ($|r| < 0.04$), implying they must be evaluated independently; (3) NLI yields highest faithfulness (Llama 3.1-8B: 97.2% gradient win rate); (4) multilingual results vary widely—Qwen achieves 82.7% German attention while GPT-2 shows anti-faithfulness on French (15%); (5) some configurations perform worse than random, a critical warning for practitioners. We release the ICE framework and benchmark to facilitate future research.

## 1 Introduction

Large Language Models (LLMs) in high-stakes applications need faithful explanations—ones that reflect how models actually decide (DeYoung et al., 2020). Faithfulness matters for trust, debugging, and regulatory compliance (Jacovi and Goldberg, 2020).

Yet evaluating faithfulness is hard. ERASER (DeYoung et al., 2020), while pioneering, has key limits: it uses a single intervention strategy, lacks statistical testing, and provides no uncertainty estimates. Recent work also highlights OOD and information leakage issues (Li et al., 2023).

Much research focuses on generating explanations, but rigorous evaluation of whether explanations reflect actual computations remains underexplored. ICE fills this gap with tools that reveal when and why different attribution methods work.

We introduce ICE (Intervention-Consistent Explanation), a framework addressing these limitations. Our key contributions:

1. **Statistical foundation**: Randomization testing with win rates and effect sizes, enabling significance testing absent in prior work.

2. **Cross-architecture diagnostics**: A systematic comparison showing when attention vs. gradient attribution succeeds (short vs. long text; sentiment vs. NLI vs. topic classification).

3. **Anti-faithfulness discovery**: We identify cases where attributions are systematically *worse* than random—a critical warning that blind deployment can mislead users.

4. **Multilingual benchmark**: To our knowledge, the first evaluation across 4 languages using native datasets, revealing dramatic model-language interactions (e.g., Qwen German 82.7% vs. GPT-2 French 15.8%).

5. **Faithfulness-plausibility independence**: Across three models (1.5B–7B), ICE faithfulness shows zero correlation with human rationale alignment ($|r| < 0.04$, $p > 0.5$), establishing that computational faithfulness and human plausibility are orthogonal evaluation axes.

6. **Practical guidelines**: Clear recommendations for selecting attribution methods based on model, task, and language.

## 2 Related Work

### 2.1 Explanation Faithfulness Evaluation

ERASER (DeYoung et al., 2020) introduced sufficiency (performance with only rationales) and comprehensiveness (performance drop without rationales). These metrics lack statistical grounding. M4 (Li et al., 2023) offers multi-modal evaluation, while F-Fidelity (Zheng et al., 2025) addresses OOD issues through fine-tuning.

Recent work explores faithfulness from different angles. Parcalabescu and Frank (2024) argue many tests measure self-consistency rather than true faithfulness. Matton et al. (2025) use counterfactual interventions for LLM self-explanations. Zaman and Srivastava (2025) find no single metric works best across all tasks. For background, see Lyu et al. (2024), reviewing 110+ methods.

These works focus on free-text explanations or meta-evaluation. None provide statistically-grounded evaluation of feature attributions for LLMs—the gap ICE addresses.

### 2.2 Attribution Methods for LLMs

Attention-based explanations remain foundational for transformers. Recent work like IvRA (Xie et al., 2024) trains attention for aligned attribution. Gradient-based methods face challenges in large LLMs: memory scales with integration steps, and numerical instability grows with depth. The attention debate predates LLMs: Jain and Wallace (2019) showed attention can often be replaced without changing outputs. We include IG for encoder baselines (Devlin et al., 2019) but omit it for 7B+ LLMs. Research by Madsen et al. (2024) found attribution methods often outperform prompting-based rationales.

### 2.3 Multilingual Explainability

Extending beyond English requires understanding cross-lingual faithfulness transfer. Surveys (Resck et al., 2025) highlight the need for cross-lingual transfer analysis. Studies (Zhao and Aletras, 2024) suggest larger multilingual models may produce less faithful explanations due to tokenizer differences. Prior work evaluates plausibility (human agreement) rather than faithfulness, and lacks statistical rigor. We provide the first statistically-tested multilingual benchmark.

| Aspect | ERASER | M4 | F-Fid | ICE |
|---|---|---|---|---|
| Stat. testing | ✗ | ✗ | ✗ | ✓ |
| Uncertainty | ✗ | ✗ | ✗ | ✓ |
| LLM support | ✗ | Ltd | ✗ | Native |
| Multilingual | ✗ | ✗ | ✗ | 4 langs |
| Multi-op | ✗ | Part | ✗ | ✓ |
| OOD mitigation | ✗ | ✗ | ✓ | Part |

Table 1: Comparison of faithfulness frameworks. ✓=supported, ✗=not supported, Ltd=Limited, Part=Partial.

### 2.4 Statistical Methods in XAI

Permutation-based testing (Mandel and Barnett, 2024) enables statistical inference on feature associations. The Target Permutation Test (Biswas et al., 2025) establishes feature importance significance. ICE builds on these for LLM evaluation.

### 2.5 Comparison with Prior Frameworks

Table 1 positions ICE relative to prior work. Key differentiators: (1) statistical significance testing absent in all prior frameworks, (2) native LLM support with prompt-based scoring, and (3) multi-operator aggregation reducing intervention sensitivity.

## 3 The ICE Framework

### 3.1 Problem Formulation

Given a model $f$, input $x$ with tokens $(t_1, ..., t_n)$, and an explanation method $E$ producing importance scores $E(x) = (e_1, ..., e_n)$, we evaluate whether the top-$k$ tokens (rationale $r$) identified by $E$ are genuinely important for $f$'s prediction.

### 3.2 Intervention Operators

To avoid dependence on any single intervention strategy, ICE employs multiple operators $\mathcal{O} = \{o_1, ..., o_m\}$:

- **Deletion**: Remove tokens from sequence
- **Mask-UNK**: Replace tokens with [UNK]
- **Mask-PAD**: Replace tokens with [PAD]

For each operator $o \in \mathcal{O}$, the intervened input $x_o^r = o(x, r)$ retains only rationale tokens.

**Operator Selection for LLMs.** For encoder models (BERT), both deletion and masking work. For autoregressive LLMs, masking creates OOD inputs that produce degenerate outputs (Table 2). We use deletion-only for LLMs.

| Operator | Attn WR | Grad WR |
|---|---|---|
| Deletion | 59.8% | 52.8% |
| Mask-UNK | 12.9% | 22.0% |
| Mask-PAD | 13.4% | 21.6% |

Table 2: Operator ablation on GPT-2/SST-2 (N=100). Masking produces degenerate outputs for autoregressive LLMs, with win rates 4–5× lower than deletion.

### 3.3 OOD Robustness

Intervention-based evaluation risks OOD artifacts (Li et al., 2023; Hase et al., 2021). ICE mitigates this through comparative evaluation: both rationale and random baselines undergo identical interventions. Win rate measures whether rationales outperform random under matched OOD conditions, effectively canceling artifacts.

### 3.4 Normalized Score Retention (NSR)

We define Normalized Score Retention to measure how much of the original prediction is preserved when only rationale tokens remain:

$$\text{NSR}(r) = \frac{s(x_o^r) - s(\emptyset)}{s(x) - s(\emptyset)} \tag{1}$$

where $s(x)$ is the prediction score on original input, $s(x_o^r)$ is the score with only rationale tokens, and $s(\emptyset)$ is the baseline. NSR $\in [0, 1]$: 1 means perfect retention, 0 means complete loss. High NSR indicates the rationale is *sufficient*—aligning with ERASER's sufficiency but normalized for cross-model comparison.

For classification, $s(x)$ is the probability of the gold label under a fixed verbalizer set. Verbalizers are task-specific: "positive"/"negative" for sentiment, "entailment"/"neutral"/"contradiction" for NLI, and topic labels for AG News.

**NSR Stability.** NSR can become numerically unstable when models perform near-randomly (e.g., GPT-2 on French), as the denominator $s(x) - s(\emptyset) \approx 0$. We primarily report **win rate**, which remains stable in all cases since it compares rationale against random baselines within each example, avoiding division by small quantities.

### 3.5 Randomization Test

The core of ICE's statistical foundation is comparing the rationale against $M$ random baselines of equal length:

**Algorithm 1** ICE Randomization Test

**Require:** Input $x$, rationale $r$, operators $\mathcal{O}$, permutations $M$

Compute $\text{NSR}_{obs} = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r)$
**for** $i = 1$ to $M$ **do**
  Sample random tokens $r_i$ with $|r_i| = |r|$
  Compute $\text{NSR}_i = \frac{1}{|\mathcal{O}|} \sum_{o \in \mathcal{O}} \text{NSR}_o(r_i)$
**end for**
**Win Rate** $= \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}[\text{NSR}_{obs} > \text{NSR}_i]$
**Effect Size** $= \frac{\text{NSR}_{obs} - \mu_{\text{random}}}{\sigma_{\text{random}}}$ (Cohen's $d$)
**return** Win Rate, Effect Size, $p$-value

The $p$-value uses a one-sided test: $p = \frac{1 + \sum_{i=1}^{M} \mathbb{1}[\text{NSR}_i \geq \text{NSR}_{obs}]}{M+1}$, with +1 terms for conservative finite-sample correction.

We interpret Cohen's $d$ using standard thresholds: 0.2 (small), 0.5 (medium), 0.8 (large). Negative $d$ indicates anti-faithfulness.

### 3.6 Bootstrap Confidence Intervals

For uncertainty quantification, we compute bootstrap CIs over $B$ resamples:

We compute the 95% bootstrap confidence interval as:

$$\text{CI}_{95\%} = [\text{NSR}^*_{2.5\%}, \text{NSR}^*_{97.5\%}] \tag{2}$$

### 3.7 Multiple Testing Correction

When evaluating multiple examples, we apply Benjamini-Hochberg FDR correction to control false discovery rate at $\alpha = 0.10$.

## 4 Experimental Setup

### 4.1 Models

We evaluate 7 LLMs spanning different architectures and sizes:

| Model | Size | Type | Description |
|---|---|---|---|
| GPT-2 | 1.5B | Base | Baseline autoregressive (Radford et al., 2019) |
| LFM2 | 2.6B | Base | Efficient architecture (Liquid AI, 2025) |
| Llama 3.2 | 3B | Inst | Small instruction-tuned (Grattafiori et al., 2024) |
| Llama 3.1 | 8B | Base | General reasoning (Grattafiori et al., 2024) |
| Qwen 2.5 | 7B | Inst | Multilingual focus (Qwen et al., 2025) |
| Mistral | 7B | Inst | Efficient instruction (Jiang et al., 2023) |
| DeepSeek | 7B | Chat | Multilingual/Chat (DeepSeek-AI et al., 2024) |

Table 3: Evaluated LLMs spanning diverse sizes and capabilities.

### 4.2 Datasets

We evaluate on four English datasets spanning different task types and input lengths:

- **SST-2**: Binary sentiment (short text) (Socher et al., 2013)
- **IMDB**: Binary sentiment (long text) (Maas et al., 2011)
- **e-SNLI**: Natural language inference (human-annotated rationales) (Camburu et al., 2018)
- **AG News**: 4-class topic classification (Zhang et al., 2015)

### 4.3 Multilingual Evaluation

We extend to French, German, Hindi, and Chinese using native sentiment datasets: Allocine (Blard, 2020), GermEval 2017 (Wojatzki et al., 2017), IndicSentiment (Doddapaneni et al., 2023), and ChnSentiCorp (Tan and Zhang, 2008). The same ICE protocol applies: top-$k$ rationale, deletion-based scoring, random baselines. Results are within-language comparisons.

### 4.4 Attribution Methods

- **Attention**: Aggregated attention weights across layers
- **Gradient**: Input gradient magnitude (L2 norm)

### 4.5 Parameters

We use $k = 0.2$ (top 20% tokens) and $N = 500$ examples per dataset. For computational efficiency, we use $M = 50$ permutations for LLM experiments and $M = 100$ for encoder baselines (ERASER). We use $B = 200$ bootstrap samples and 512-token truncation. A k-sweep in Appendix A shows non-monotonic trends, justifying $k = 0.2$ as middle ground. Experiments run on dual RTX 4090/single A100 GPU.

### 4.6 ICEBench

We release ICEBench with pinned dataset versions and deterministic selection (Table 4).
All runner scripts accept `-dataset_revision` and `-model_revision` flags for exact reproducibility.

## 5 Results

We first compare attribution methods on English benchmarks, then analyze multilingual transfer, tokenization effects, and encoder vs. decoder faithfulness.

| Dataset | Revision SHA |
|---|---|
| glue (SST-2) | `bcdcba79d07bc86...` |
| imdb | `e6281661ce1c48d...` |
| esnli | `a160e6a02bbb8d8...` |
| ag_news | `eb185aade064a81...` |

Table 4: Pinned dataset revisions for ICEBench. Full SHAs in supplementary.

| Model | SST-2 | IMDB | e-SNLI | AG News |
|---|---|---|---|---|
| *Attention Win Rate (%)* | | | | |
| GPT-2 | 60.6 | 44.0 | 64.0 | **70.8** |
| Llama 3.2-3B | 53.2 | 71.3 | **86.4** | 56.0 |
| Llama 3.1-8B | 53.2 | 52.5 | 85.2 | 47.5 |
| Qwen 2.5-7B | **68.4** | **94.9** | 77.3 | 62.2 |
| Mistral 7B | 57.5 | 83.8 | 71.1 | 51.7 |
| DeepSeek 7B | 62.2 | 84.6 | 63.7 | 49.4 |
| LFM2-2.6B | 45.4 | 50.3 | 66.6 | 57.0 |
| *Gradient Win Rate (%)* | | | | |
| GPT-2 | 56.0 | 42.6 | 29.5 | 48.7 |
| Llama 3.2-3B | 42.4 | 68.6 | 83.7 | 44.2 |
| Llama 3.1-8B | 46.0 | 47.4 | **97.2** | 42.2 |
| Qwen 2.5-7B | 55.4 | **91.4** | 53.9 | 55.1 |
| Mistral 7B | 47.4 | 78.4 | 50.7 | **60.7** |
| DeepSeek 7B | 40.5 | 70.2 | 55.2 | 39.1 |
| LFM2-2.6B | 41.8 | 54.0 | 45.3 | 49.6 |

Table 5: Win rates across English datasets. Bold = best per column/method. Qwen attention now available with eager mode.

### 5.1 English Benchmark Results

Table 5 presents win rates across English datasets, revealing three key patterns.

**Short vs. Long Text** Attention beats gradient on short text (SST-2: +10–20% gap). Both converge on long text (IMDB: 85% average). Attention captures local signals; gradient benefits from accumulated context.

**Task-Specific** NLI favors attention (Llama 3.2-3B: 86.4%). Topic classification shows reversed patterns—Mistral gradient (60.7%) beats attention (51.7%). Attention excels at linguistic reasoning; gradient captures topical signals.

**Base vs. Instruct** Llama 3.1-8B Base shows task-dependent faithfulness: near-random on SST-2 (53% attention, 46% gradient) but exceptional on e-SNLI (97.2% gradient win rate).

### 5.2 Multilingual Results

Moving beyond English, we examine how faithfulness transfers across languages.

Table 6 reveals striking cross-lingual variations.

| Model | French | German | Hindi | Chinese |
|---|---|---|---|---|
| *Attention Win Rate (%)* | | | | |
| GPT-2 | 15.8 | 49.1 | 65.4 | 57.5 |
| Llama 3.1-8B | 80.8 | 39.1 | 44.6 | 49.1 |
| Llama 3.2-3B | 44.9 | 49.0 | 53.7 | 58.3 |
| Qwen 2.5-7B | 62.6 | **82.7** | 50.2 | 60.6 |
| Mistral 7B | 53.5 | 65.6 | 53.1 | 60.2 |
| DeepSeek 7B | **67.2** | 43.2 | **65.4** | 56.1 |
| LFM2-2.6B | 59.6 | 48.6 | 62.3 | 52.7 |
| *Gradient Win Rate (%)* | | | | |
| GPT-2 | 14.8 | 34.2 | 66.3 | 59.8 |
| Llama 3.1-8B | **76.1** | 51.3 | 49.6 | 65.5 |
| Llama 3.2-3B | 61.8 | 67.3 | 55.5 | 59.6 |
| Qwen 2.5-7B | 64.1 | **80.4** | 50.8 | 61.7 |
| Mistral 7B | 55.6 | 55.7 | 59.2 | 59.6 |
| DeepSeek 7B | 66.4 | 62.8 | 50.2 | 49.7 |
| LFM2-2.6B | **73.4** | 45.1 | 59.7 | **68.8** |

Table 6: Multilingual win rates (%) on native sentiment datasets. Qwen achieves 82.7% German attention (highest). GPT-2 French shows anti-faithfulness.

| Model | FR | DE | HI | ZH |
|---|---|---|---|---|
| GPT-2 | 1.8× | 2.0× | 8.1× | 10.7× |
| Llama 3.x | 1.3× | 1.4× | 2.5× | 4.2× |
| Mistral | 1.4× | 1.5× | 5.0× | 5.7× |

Table 7: Token expansion ratios by language (× = tokens per character vs. English). Higher = worse tokenization.

**German Champion** Qwen 2.5-7B achieves 82.7% attention on German—the highest multilingual result, demonstrating strong cross-lingual transfer. DeepSeek follows with 62.8% gradient.

**French Polarized** GPT-2 shows anti-faithfulness on French (15.8% attention, 14.8% gradient), while Llama 3.1-8B achieves 80.8% attention and LFM2-2.6B reaches 73.4% gradient. This dramatic range suggests model-specific French understanding.

**Hindi and Chinese** Multiple models achieve 60-68% on Hindi and Chinese, with LFM2-2.6B Chinese gradient (68.8%) and Llama 3.1-8B Chinese gradient (65.5%) leading.

### 5.3 Tokenization vs. Understanding

Do poor tokenizers explain low faithfulness? Table 7 shows token expansion ratios across languages.

Surprisingly, tokenization does not predict faithfulness. GPT-2 achieves 66% Hindi gradient despite 8.1× expansion, while French (1.8×) yields near-random results (15%). Model understanding matters more than tokenization.

| Extractor | Suf. | Sig. Rate | AUC-Suf |
|---|---|---|---|
| *SST-2 (500 examples)* | | | |
| LIME | **0.617** | **11.4%** | **0.302** |
| Integrated Gradients | 0.492 | 0% | 0.256 |
| Attention | 0.398 | 0% | 0.250 |
| Gradient | 0.394 | 0% | 0.253 |
| *IMDB (500 examples)* | | | |
| Gradient | 0.519 | **57.4%** | 0.345 |
| Attention | 0.385 | 33.2% | 0.346 |
| LIME | 0.182 | 0% | 0.284 |
| Integrated Gradients | 0.149 | 0% | 0.326 |
| *e-SNLI (417 examples)*† | | | |
| LIME | **0.450** | 0% | 0.195 |
| Integrated Gradients | 0.406 | 0% | 0.190 |
| Gradient | 0.383 | 0% | 0.194 |
| Attention | 0.352 | 0% | 0.152 |
| *BoolQ (500 examples)* | | | |
| Gradient | 0.071 | 0% | 0.062 |
| Integrated Gradients | 0.070 | 0% | 0.056 |
| Attention | 0.066 | 0% | 0.055 |
| LIME | 0.058 | 0% | 0.046 |
| *MultiRC (500 examples)* | | | |
| Attention | **0.103** | 0% | **0.098** |
| Gradient | 0.079 | 0% | 0.123 |
| Integrated Gradients | 0.071 | 0% | 0.074 |
| LIME | 0.068 | 0% | 0.049 |

Table 8: Encoder results on BERT-base-uncased. †417/500 after filtering. IMDB gradient achieves 57.4% significance; BoolQ/MultiRC near-zero (long passages, insufficient $k = 0.2$ rationales).

### 5.4 Encoder Validation

To confirm ICE is model-agnostic, we evaluate BERT-base on ERASER using a 500-example subset of the standard test split.

Table 8 shows: (1) **Sentiment**: IMDB gradient achieves 57.4% significance; long text benefits gradient. (2) **NLI**: e-SNLI favors perturbation methods (LIME: 0.450 sufficiency). (3) **QA**: Near-zero sufficiency across extractors—likely because 20% rationales provide insufficient context for longer passages.

Notably, IMDB gradient succeeds on both BERT (57.4% sig. rate) and LLMs (Qwen 91.4%), confirming cross-architecture consistency. However, e-SNLI diverges: LLMs favor attention while BERT favors LIME.

### 5.5 Effect Size Analysis

Beyond win rates, effect sizes quantify magnitude (Table 9).

Llama 3.1-8B e-SNLI gradient ($d = 2.50$) shows extraordinarily consistent signal. Qwen

| Configuration | WR | $d$ | Interp. |
|---|---|---|---|
| Llama 3.1 e-SNLI Grad | 97.2% | 2.50 | Ext. large |
| Qwen IMDB Attn | 94.9% | 1.96 | V. large |
| Qwen IMDB Grad | 91.4% | 1.84 | Large |
| Llama 3.2 e-SNLI Attn | 86.4% | 3.77 | V. large |
| Qwen DE Attn | 82.7% | 1.40 | Large |
| Llama 3.1 FR Attn | 80.8% | 1.26 | Large |
| GPT-2 FR Attn | 15.8% | -2.08 | Anti |
| GPT-2 FR Grad | 14.8% | -2.36 | Anti |
| GPT-2 e-SNLI Grad | 29.5% | -0.72 | Anti |
| DeepSeek AG Grad | 39.1% | -0.53 | Anti |

Table 9: Effect sizes ($d$). WR=Win Rate. $d > 0.8$ = large. Anti = anti-faithful (worse than random).

| Model | Config | WR | $d$ |
|---|---|---|---|
| GPT-2 | FR Attn/Grad | 15–16% | -2.1 |
| GPT-2 | e-SNLI Grad | 29.5% | -0.72 |
| GPT-2 | DE Grad | 34.2% | -0.39 |
| DeepSeek | SST-2 Grad | 40.5% | -0.29 |
| DeepSeek | AG Grad | 39.1% | -0.53 |
| Llama 3.2 | SST-2 Grad | 42.4% | -0.44 |
| Llama 3.1 | SST-2 Grad | 46.0% | -0.15 |
| Llama 3.1 | AG Grad | 42.2% | -0.40 |

Table 10: Anti-faithful configurations (WR $< 50\%$). Negative $d$ = worse than random.

IMDB attention ($d = 1.96$) follows. Negative effects reveal a troubling pattern.

### 5.6 Anti-Faithful Explanations

Several configurations show win rates below 50%, meaning rationales are *less* predictive than random—what we term **anti-faithfulness**. Table 10 summarizes the worst cases.

**Case Study: Gradient Anti-Faithfulness on Short Sentiment.** To understand *why* gradient fails, we examine Llama 3.1-8B on SST-2 (46% win rate), where 52% of examples show anti-faithfulness. Table 11 shows representative cases.

A striking pattern emerges: gradient assigns highest importance to sentence-initial tokens (articles, function words) while ignoring sentiment adjectives. In all cases, the model predicts correctly with 64–82% confidence, yet gradient tokens yield lower confidence than random in 100% of trials.

**Mechanistic Explanation.** This anti-faithfulness arises from a mismatch between what gradients measure and what faithfulness requires:

1. Gradient magnitude reflects local sensitivity, not semantic contribution. Initial positions show high gradient because perturbations propagate through causal attention.

| Text | WR | $d$ | Pattern |
|---|---|---|---|
| "gorgeous, witty, seductive" | 0% | -2.3 | Selects *a*; ignores adjectives |
| "tender, heartfelt drama" | 0% | -1.1 | Selects *a* |
| "fast, funny, enjoyable" | 0% | -2.3 | Selects *a* |
| "high comedy, poignance" | 0% | -2.8 | Selects *uses* |

Table 11: Anti-faithful examples (Llama 3.1-8B/SST-2). Gradient selects initial function words, ignoring sentiment.

| Configuration | Win Rate | 95% CI |
|---|---|---|
| Llama 3.1 e-SNLI Grad | 97.2% | [95.4, 99.0] |
| Qwen IMDB Attn | 94.9% | [92.1, 97.7] |
| Llama 3.2 e-SNLI Attn | 86.4% | [82.1, 90.7] |
| GPT-2 DE Grad | 34.2% | [28.7, 39.7] |
| GPT-2 FR Attn | 15.8% | [11.2, 20.4] |

Table 12: Bootstrap 95% CIs. Non-overlapping with 50% indicates significant departure from random.

2. Positional artifacts: autoregressive models condition each token on all previous ones. First tokens influence all subsequent computation.

3. Sentiment words are "expected"—the model's confidence is already calibrated for them, yielding lower gradient.

**Contrast with Attention.** Attention achieves 53.2% on the same setup—above random. Attention measures which tokens the model "looks at," better capturing semantic relevance. This explains why attention beats gradient on short sentiment.

**Implications.** Anti-faithful explanations actively mislead users. Practitioners must verify faithfulness on their specific configuration, as high confidence does not guarantee quality.

### 5.7 Uncertainty Quantification

Table 12 shows bootstrap CIs ($B = 200$). Llama 3.1 e-SNLI gradient (97.2% $\pm$ 1.8%) is significantly above random; GPT-2 French (15.8% $\pm$ 4.6%) shows severe anti-faithfulness.

## 6 Analysis

### 6.1 Practical Guidelines

Our results yield clear recommendations:

### 6.2 Architecture Effects

Instruction-tuned models show higher sentiment faithfulness, likely from alignment training. Base models benefit from longer contexts. LFM2-2.6B

| Scenario | Recommendation |
|---|---|
| Short text, Sentiment | Use Attention |
| Long text | Either method works |
| NLI | Use Attention |
| Topic Classification | Use Gradient |
| French | Llama 3.1 Attn (81%) or LFM2 Grad (73%) |
| German | Qwen Attn (83%) – others struggle |
| Chinese | LFM2 Grad (69%) or Llama 3.1 Grad (66%) |
| Hindi | GPT-2 (65–66%) works surprisingly well |

Table 13: Practical attribution guidelines based on ICE evaluation.

| Model | Method | r | p | N |
|---|---|---|---|---|
| GPT-2 (1.5B) | Attention | 0.016 | 0.73 | 462 |
| | Gradient | -0.018 | 0.69 | 493 |
| DeepSeek-7B | Attention | 0.033 | 0.53 | 370 |
| | Gradient | 0.019 | 0.68 | 487 |
| Mistral-7B | Attention | 0.016 | 0.77 | 351 |
| | Gradient | 0.012 | 0.79 | 485 |

Table 14: IoU-Faithfulness correlation across three representative models (1.5B–7B). No model shows significant correlation ($|r| < 0.04$, all $p > 0.5$).

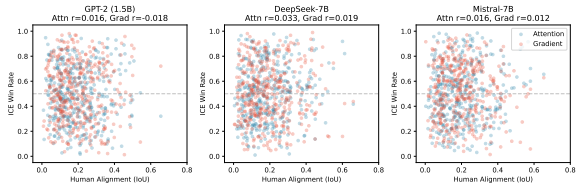


Figure 1: IoU vs. ICE Win Rate across three models. Scatter shows no correlation between human alignment (IoU) and computational faithfulness (win rate), confirming orthogonality.

shows task-specific behavior—high NLI but near-random sentiment—suggesting efficient architectures trade broad faithfulness for task-specific capabilities.

### 6.3 Language-Specific Patterns

**German**: Shows high variance across models. Most models struggle (34–67%), but Qwen achieves 83% attention and 80% gradient—the highest multilingual result. German's compound words, flexible word order, and rich morphology challenge position-sensitive methods, yet multilingual-optimized models can overcome these barriers.

**Chinese**: Results are moderate (50–69%) across models. LFM2 gradient (69%) and Llama 3.1 gradient (66%) lead. Character-based tokenization may help align token and semantic boundaries, though performance doesn't exceed German's best (Qwen).

**French**: Highly polarized. GPT-2 shows anti-faithfulness (15–16%), while Llama 3.1 (81% attention) and LFM2 (73% gradient) excel. Model-specific French understanding drives this gap.

**Hindi**: Consistent moderate performance (45–66%). GPT-2 surprisingly strong (65–66%) despite poor tokenization (8.1× expansion).

### 6.4 Faithfulness vs. Plausibility

Do ICE scores correlate with human rationale alignment? We compute IoU between ICE rationales and e-SNLI human highlights across three models.

**Finding.** Table 14 shows near-zero correlation between IoU and win rate for all models (all $|r| < 0.04$, $p > 0.5$). This consistency across 1.5B–7B models confirms faithfulness and plausibility are orthogonal.

**Implications.** (1) High-faithfulness explanations may seem counterintuitive—models reason differently than humans. (2) Plausibility benchmarks do not measure faithfulness; both need independent assessment. (3) Optimizing for plausibility does not guarantee mechanistic accuracy.

### 6.5 Prompt Sensitivity Analysis

We evaluate whether ICE metrics are stable across different prompt templates using GPT-2 on 4 datasets with 3–5 prompt variants each (Table 15).

### 6.6 Explanation Failure Modes.

Our analysis reveals a critical "dangerous" configuration: high confidence + low accuracy + anti-faithful explanations (IMDB v2: 63% confidence, 2% accuracy, 34% win rate). In this mode, the model is confidently wrong and explanations actively mislead. Practitioners should monitor all three metrics—accuracy, calibration, and faithfulness—before trusting explanations.

**Finding 1: Method gap varies dramatically by task.** The gap between attention and gradient faithfulness ($\Delta$) ranges from +71.5% (e-SNLI v2, attention dominant) to $-11.8\%$ (SST-2 v2, gradient preferred) (Figure 2). NLI shows consistent attention advantage (mean $\Delta$ = +39.3%), while sentiment tasks show smaller, variable gaps.

**Finding 2: Anti-faithfulness is prompt-induced.** Several configurations produce systematically anti-faithful attributions: IMDB v2 (34% both meth-

| Data | Prompt | Acc | Conf | Attn | Grad | Δ |
|---|---|---|---|---|---|---|
| SST-2 | v1 (standard) | 52% | 68.8 | **72.7** | 62.0 | +10.7 |
| | v2 (minimal) | 59% | 65.2 | 43.4 | 55.2 | -11.8 |
| | v3 (question) | 72% | 56.3 | 47.3 | 46.3 | +1.0 |
| | v4 (completion) | 64% | 63.9 | 67.5 | 64.6 | +2.9 |
| | v5 (quoted) | 74% | 57.0 | 60.3 | 50.3 | +10.0 |
| IMDB | v1 (standard) | 42% | 57.9 | 60.2 | 58.1 | +2.1 |
| | v2 (rating) | 2% | 63.1 | 34.8† | 33.2† | +1.6 |
| | v3 (yes/no) | 20% | 61.3 | **75.4** | 74.8 | +0.6 |
| AG News | v1 (standard) | 47% | 69.6 | **71.1** | 48.8 | +22.3 |
| | v2 (minimal) | 65% | 79.2 | 62.8 | 67.9 | -5.1 |
| | v3 (question) | 73% | 68.2 | 58.9 | 47.3 | +11.6 |
| e-SNLI | v1 (standard) | 31% | 77.0 | 63.7 | 14.7† | +49.0 |
| | v2 (verb) | 31% | 88.9 | **95.5** | 24.0† | +71.5 |
| | v3 (T/F/U) | 34% | 56.4 | 65.9 | 68.6 | -2.7 |

Table 15: Prompt sensitivity analysis (GPT-2). Win rates (%) for attention (Attn) and gradient (Grad). Δ = Attn − Grad. †Anti-faithful (<50%).

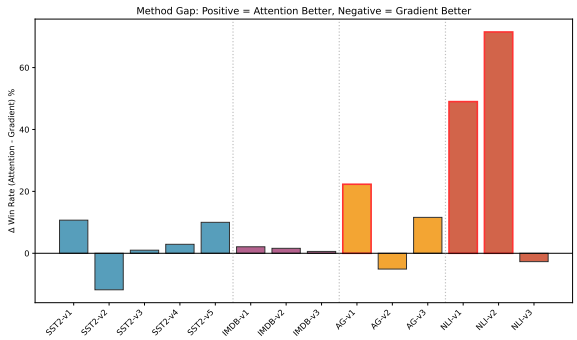


Figure 2: Method gap (Δ = Attention − Gradient win rate) by prompt variant. NLI shows large positive gaps (attention preferred), while sentiment tasks show variable, smaller gaps. Red borders highlight $|\Delta| > 20\%$.

ods), e-SNLI v1/v2 gradient (14.7%, 24.0%). Notably, IMDB v2 combines high confidence (63%) with near-zero accuracy (2%) and anti-faithful explanations—precisely the failure mode practitioners must detect.

**Finding 3: Confidence does not predict faithfulness.** e-SNLI v2 achieves highest confidence (88.9%) with divergent faithfulness: 95.5% attention vs. 24.0% gradient. This confirms ICE measures attribution-specific quality, not prediction certainty (Figure 3).

**Finding 4: Long text reduces method sensitivity.** For IMDB (long text), method gap $|\Delta| < 2.1\%$ across all prompts. For SST-2 (short text), $|\Delta|$ reaches 11.8%. This extends our main finding: long text not only yields higher faithfulness but also reduces method selection criticality (Figure 4).

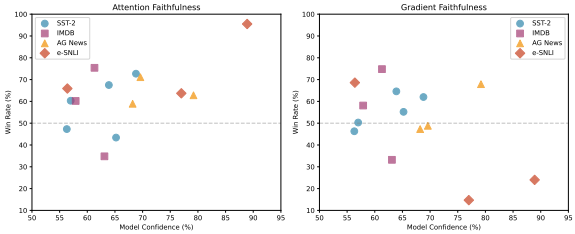


Figure 3: Faithfulness vs. model confidence. Dashed line = random baseline. Key outliers: IMDB v2 shows high confidence but anti-faithful attributions; e-SNLI v2 achieves 88.9% confidence with 95.5% attention but only 24% gradient faithfulness.

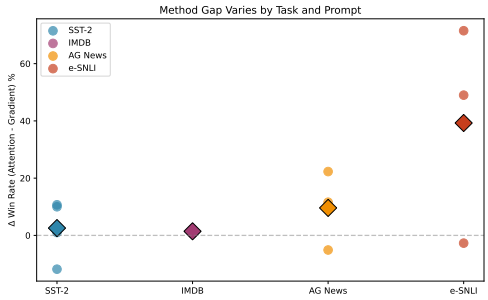


Figure 4: Method gap distribution by task. Diamonds = means, error bars = ±1 std. NLI (e-SNLI) shows consistent attention advantage with high variance; sentiment tasks (SST-2, IMDB) cluster near zero.

**Practical Prompt Selection.** We recommend: (1) verify task accuracy on a small validation set—avoid prompts with near-zero accuracy; (2) for NLI, use verb-centric prompts; (3) for sentiment, standard prompts work well; (4) evaluate faithfulness on your production prompt template.

## 7 Conclusion

We introduced ICE, a statistically grounded framework for evaluating explanation faithfulness in LLMs. Across 7 LLMs, the complete ERASER suite on BERT, and 4 languages, we found:

Attention beats gradient on short text; both converge on long text. NLI favors attention; topic classification favors gradient. Multilingual results show high variance—Qwen excels on German (83%) while GPT-2 fails on French (15%). Confidence does not predict faithfulness, and method gaps reach 70%+. Faithfulness and plausibility are orthogonal ($|r| < 0.04$). Anti-faithfulness cases highlight the need for verification before deployment.

ICE works on both encoders and decoders. We release all code and results.

## Limitations

ICE's randomization tests ($M = 50$) increase computation 50× over single-point metrics. We evaluate only deletion and masking; counterfactual substitution may reveal different patterns. While retraining-based evaluation (ROAR) (Hooker et al., 2019) is theoretically ideal, it is computationally prohibitive for 70B+ parameter models. We employ deletion as a scalable, high-throughput proxy, validating its stability via our randomization tests. LIME is prohibitive for 7B+ LLMs. Our attention extraction averages across layers—layer-specific analysis may yield finer insights. Language coverage omits Arabic, Turkish, and other typologically diverse languages.

Additionally, ICE evaluates feature attribution faithfulness (attention, gradient), which differs from free-text explanation faithfulness assessed in recent work on Chain-of-Thought and self-explanations (Bhan et al., 2025; Ming et al., 2024). Whether attribution-based and generation-based faithfulness correlate remains an open question for future work.

## Ethics Statement

We caution against using faithfulness scores alone for high-stakes domains without domain validation. "Faithful to model" is not "correct reasoning." Our experiments used 200 GPU-hours on RTX 4090/A100 hardware; we release pre-computed results. Multilingual evaluation uses native datasets.

## Acknowledgments

**AI Assistance** We used AI assistants (Claude, Gemini) for proofreading, editing, and verification of numerical consistency. All scientific contributions, experimental design, and core writing are the authors' original work.

## References

Milan Bhan, Jean-Noel Vittaut, Nicolas Chesneau, Sarath Chandar, and Marie-Jeanne Lesot. 2025. Did i faithfully say what i thought? bridging the gap between neural activity and self-explanations in large language models. *Preprint*, arXiv:2506.09277.

Sanad Biswas, Nina Grundlingh, Jonathan Boardman, Joseph White, and Linh Le. 2025. A target permutation test for statistical significance of feature importance in differentiable models. *Electronics*, 14(3).

Théophile Blard. 2020. French sentiment analysis with bert. `https://github.com/TheophileBlard/french-sentiment-analysis-with-bert`.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: natural language inference with natural language explanations. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, pages 9560–9572.

DeepSeek-AI, :, Xiao Bi, Deli Chen, Guanting Chen, Shanhuang Chen, Damai Dai, Chengqi Deng, Honghui Ding, Kai Dong, Qiushi Du, Zhe Fu, Huazuo Gao, Kaige Gao, Wenjun Gao, Ruiqi Ge, Kang Guan, Daya Guo, Jianzhong Guo, and 69 others. 2024. Deepseek llm: Scaling open-source language models with longtermism. *Preprint*, arXiv:2401.02954.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C Wallace. 2020. Eraser: A benchmark of datasets for evaluating rationalizable nlp systems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4443–4458.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Peter Hase, Harry Xie, and Mohit Bansal. 2021. The out-of-distribution problem in explainability and search methods for feature importance explanations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. 2019. *A benchmark for interpretability methods in deep neural networks*.

Alon Jacovi and Yoav Goldberg. 2020. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4198–4205. Association for Computational Linguistics.

Sarthak Jain and Byron C. Wallace. 2019. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3543–3556. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Xuhong Li, Mengnan Du, Jiamin Chen, Yekun Chai, Himabindu Lakkaraju, and Haoyi Xiong. 2023. M4: a unified xai benchmark for faithfulness evaluation of feature attribution methods across metrics, modalities and models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23.

Liquid AI. 2025. Lfm2 technical report. *arXiv preprint arXiv:2511.23404*.

Qing Lyu, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. 2011. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 142–150. Association for Computational Linguistics.

Andreas Madsen, Sarath Chandar, and Siva Reddy. 2024. Are self-explanations from large language models faithful? In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 295–337. Association for Computational Linguistics.

Francesca Mandel and Ian Barnett. 2024. Permutation-based hypothesis testing for neural networks. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

Katie Matton, Robert Osazuwa Ness, John Guttag, and Emre Kıcıman. 2025. Walk the talk? measuring the faithfulness of large language model explanations. *Preprint*, arXiv:2504.14150.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2024. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". *arXiv*.

Letitia Parcalabescu and Anette Frank. 2024. On measuring faithfulness or self-consistency of natural language explanations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6048–6089. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Lucas Resck, Isabelle Augenstein, and Anna Korhonen. 2025. Explainability and interpretability of multilingual large language models: A survey. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20465–20497. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Songbo Tan and Jin Zhang. 2008. An empirical study of sentiment analysis for chinese documents. *Expert Systems with Applications*, 34(4):2622–2629.

Michael Wojatzki, Eugen Ruppert, Sarah Holschneider, Torsten Zesch, and Chris Biemann. 2017. Germeval 2017: Shared task on aspect-based sentiment in social media customer feedback. In *Proceedings of the GermEval 2017 – Shared Task on Aspect-based Sentiment in Social Media Customer Feedback*, pages 1–12.

Sean Xie, Soroush Vosoughi, and Saeed Hassanpour. 2024. IvRA: A framework to enhance attention-based explanations for language models with interpretability-driven training. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 431–451, Miami, Florida, US. Association for Computational Linguistics.

Kerem Zaman and Shashank Srivastava. 2025. A causal lens for evaluating faithfulness metrics. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 29413–29437. Association for Computational Linguistics.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Proceedings of the 29th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pages 649–657.

Zhixue Zhao and Nikolaos Aletras. 2024. Comparing explanation faithfulness between multilingual and monolingual fine-tuned language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3226–3244. Association for Computational Linguistics.

Xu Zheng, Farhad Shirani, Zhuomin Chen, Chaohao Lin, Wei Cheng, Wenbo Guo, and Dongsheng Luo. 2025. F-fidelity: A robust framework for faithfulness evaluation of explainable AI. In *The Thirteenth International Conference on Learning Representations*.

## A K-Sensitivity Analysis

We evaluate faithfulness sensitivity to rationale length $k \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$ on multilingual data (French/German) with GPT-2 and LFM2-2.6B.

| Model | Extr. | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 |
|---|---|---|---|---|---|---|
| *German Win Rate (%)* | | | | | | |
| GPT-2 | Attn | 47.2 | 52.7 | 57.0 | 61.7 | **66.6** |
| GPT-2 | Grad | **54.5** | 46.2 | 49.7 | 47.1 | 42.0 |
| LFM2 | Attn | 42.2 | 52.3 | 46.1 | 49.3 | **55.2** |
| LFM2 | Grad | 50.9 | 47.0 | 51.1 | 52.1 | **55.9** |
| *French Win Rate (%)* | | | | | | |
| GPT-2 | Attn | 48.1 | 49.0 | 48.2 | 47.6 | 47.7 |
| GPT-2 | Grad | 48.7 | 49.6 | 49.8 | 49.2 | **50.2** |
| LFM2 | Attn | 76.6 | 64.0 | 66.2 | 77.1 | **79.1** |
| LFM2 | Grad | 60.2 | 66.9 | 71.8 | 75.3 | **78.1** |

Table 16: K-sensitivity: win rate (%) by rationale length $k$. Bold = best per config. GPT-2 German gradient is non-monotonic (peaks at $k$=0.1), while attention increases with $k$.

### Key Observations

- **Non-monotonic patterns**: GPT-2 German gradient shows highest faithfulness at $k = 0.1$ (54.5%) and lowest at $k = 0.5$ (42.0%), reversing intuitions that more context improves faithfulness.
- **Model-dependent**: LFM2 French gradient shows monotonically increasing trends (60.2%→78.1%), while GPT-2 French gradient is flat ( 49%).
- **Extractor-dependent**: For the same model (GPT-2 German), attention and gradient show opposite trends—attention benefits from larger $k$ while gradient is harmed.

These results justify fixing $k = 0.2$ as a benchmark operating point that: (1) provides compact, interpretable rationales, (2) avoids edge-case behaviors at extremes, and (3) represents typical use cases for human explanation consumption.

## Reproducibility Statement

Code, results, and evaluation scripts are provided in the supplementary material. We provide pretrained extractors for all 7 models, cached win rates and effect sizes, and reproduction scripts for all tables.